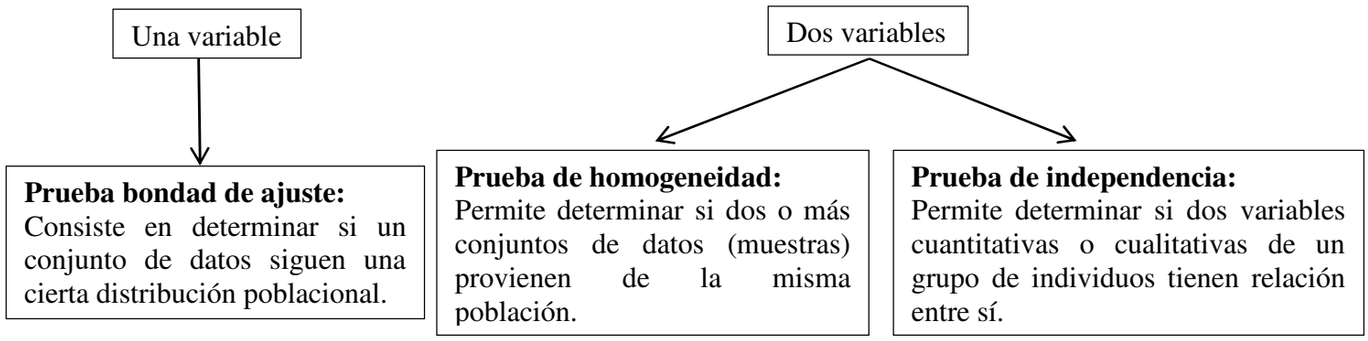
Pruebas de Chi-cuadrado



Prueba bondad de ajuste.

Sea $X \sim F_0(\theta)$, F_0 es una distribución teórica conocida con parámetro θ, y una muestra aleatoria de tamaño n de X, agrupadas en m intervalos $I_1, I_2, .., I_m$, con frecuencias observadas $n_1, n_2, .., n_m$. Entonces, para probar la hipótesis que la muestra aleatoria de X sigue una distribución teórica F_0, se debe seguir los siguientes pasos:

P1) **Plantear Hipótesis:**
H_0 : Los datos se ajustan a la distribución F_0 v/s H_1 : Los datos no se ajustan a la distribución F_0

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{m-k-1}$,

donde

n_i : Frecuencia observada en el intervalo I_i, $i = 1, .., m$.

$e_i = n \cdot P(I_i)$, $i = 1, .., m$, es la frecuencia esperada en el intervalo I_i.

k : Número de parámetros estimados de F_0.

m : Número de intervalos que agrupan los datos.

P3) **Establecer nivel de significancia:** α

P4) **Región de rechazo de H_0:** H_0 v/s $H_1 \Rightarrow R = \left(\chi^2_{1-\alpha, m-k-1}, \infty\right)$.

P5) **Decisión:** Si $J_0 \in R$, entonces H_0 se rechaza al nivel de significancia α.

P6) **Conclusión:** Se debe interpretar la decisión tomada en P5).

Ejemplo:

1) Se ha tomado una muestra aleatoria de 40 baterías y se ha registrado su tiempo de duración en años. Estos resultados se agrupan de la siguiente forma:

i	Intervalo (años)	N° de baterías (n_i)
1	1,45 - 1, 95	2
2	1, 95 - 2,45	1
3	2,45 - 2,95	4
4	2,95 - 3,45	15
5	3,45 - 3,95	10
6	3,95 - 4,45	5
7	4,45 - 4,95	3

Verificar si los tiempos de duración de las baterías producidas por el fabricante tiene una distribución normal con media 3,5 y desviación estándar 0,7. Use $\alpha = 0,05$.

Respuesta:
Sea X : "Tiempo de duración de las baterías"

P1) **Plantear Hipótesis:** $H_0: X \sim N(3,5;0.49)$ v/s $H_1: X \neq N(3,5;0.49)$

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{m-k-1}$

i	a_i	b_i	n_i	$P(X \leq b_i)$	$p_i = P(a_i < X < b_i)$	$e_i = n \cdot pi$
1	1,45	1,95	2	0,01340	0,0134	0,5362
2	1,95	2,45	1	0,06681	0,0534	2,1361
3	2,45	2,95	4	0,21602	0,1492	5,9684
4	2,95	3,45	15	0,47153	0,2555	10,2204
5	3,45	3,95	10	0,73984	0,2683	10,7325
6	3,95	4,45	5	0,91263	0,1728	6,9116
7	4,45	4,95	3		0,0874	3,4947
			$n = 40$		$\sum_{i=1}^{7} p_i \approx 1$	$\sum_{i=1}^{7} e_i \approx 40$

$$P(X \leq b_1) = P(X \leq 1,95) = \Phi\left(\frac{1,95-3,5}{0,7}\right) = 0,0134$$
$$P(X \leq b_2) = P(X \leq 2,45) = \Phi\left(\frac{2,45-3,5}{0,7}\right) = 0,0668$$
$$P(X \leq b_3) = P(X \leq 2,95) = \Phi\left(\frac{2,95-3,5}{0,7}\right) = 0,2160$$
$$P(X \leq b_4) = P(X \leq 3,45) = \Phi\left(\frac{3,45-3,5}{0,7}\right) = 0,4715$$
$$P(X \leq b_5) = P(X \leq 3,95) = \Phi\left(\frac{3,95-3,5}{0,7}\right) = 0,7398$$
$$P(X \leq b_6) = P(X \leq 4,45) = \Phi\left(\frac{4,45-3,5}{0,7}\right) = 0,9126$$
$$p_7 = P(X > b_6) = 1 - P(X \leq b_6) = 0,0874$$

Es necesario que se cumpla $e_i \geq 5$, I_i, $i = 1,..,7$.

i	n_i	e_i
1	7	8,6407
2	15	10,2204
3	10	10,7325
4	8	10,4063
	40	$\sum_{i=1}^{7} e_i \approx 40$

$$J_0 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i} = \frac{(7-8,6407)^2}{8,6407} + \frac{(15-10,2204)^2}{10,2204} + \frac{(10-10,7325)^2}{10,7325} + \frac{(8-10,4063)^2}{10,4063} = 3,1531$$

P3) **Establecer nivel de significancia:** $\alpha = 0,05$.

P4) **Región de rechazo de H_0:**

$m = 4, \quad k = 0$

$$H_0 \text{ v/s } H_1 \Rightarrow R = \left(\chi^2_{1-\alpha,m-k-1}, \infty\right) = \left(\chi^2_{0,95,3}, \infty\right) = \left(7,815; \infty\right).$$

P5) **Decisión:** Si $J_0 = 3{,}1531 \notin R = (7{,}815;\infty)$, entonces H_0 no se rechaza.

P6) **Conclusión:** Con un 95% de confianza se puede afirmar que los tiempos de duración de las baterías producidas por el fabricante tiene una distribución normal con media 3,5 y desviación estándar 0,7.

Ejemplo:

2) Para analizar el número de defectos por artículo de una fábrica, se toma una muestra aleatoria de tamaño n=60. Obteniéndose los siguientes resultados:

i	N° defectos	N° artículos (n_i)
1	0	32
2	1	15
3	2	9
4	3	4
	Total	60

¿El número de defectos por artículo tienen una distribución de Poisson?. Use $\alpha = 0{,}05$.

Respuesta:
Sea X : "Número de defectos por artículo".

P1) **Plantear Hipótesis:** $H_0: X \sim Poisson(\lambda)$ v/s $H_1: X \neq Poisson(\lambda)$

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{m-k-1}$

$$P(X = x) = \frac{e^{-\hat{\lambda}} \hat{\lambda}^x}{x!}, \quad x = 0,1,2,...$$

$$\hat{\lambda} = \overline{x} = \frac{0 \cdot 32 + 1 \cdot 15 + 2 \cdot 9 + 3 \cdot 4}{60} = 0{,}75$$

i	x_i	n_i	$p_i = P(X = x_i)$	$e_i = n \cdot pi$
1	0	32	0,4724	28,3420
2	1	15	0,3543	21,2565
3	2	9	0,1329	7,9712
4	3	4	0,0405	2,4303
		$n = 60$	$\sum_{i=1}^{4} p_i \approx 1$	$\sum_{i=1}^{4} e_i \approx 60$

$$p_1 = P(X = x_1) = P(X = 0) = \frac{e^{-0{,}75}(0{,}75)^0}{0!} = 0{,}4724$$

$$p_2 = P(X = x_2) = P(X = 1) = \frac{e^{-0{,}75}(0{,}75)^1}{1!} = 0{,}3543$$

$$p_3 = P(X = x_3) = P(X = 2) = \frac{e^{-0{,}75}(0{,}75)^2}{2!} = 0{,}1329$$

$$p_4 = P(X \geq x_4) = 1 - P(X < x_4) = 1 - [p_1 + p_2 + p_3] = 0{,}0405$$

Es necesario, que se cumpla $e_i \geq 5$, I_i, $i = 1,..,4$. Luego,

i	n_i	e_i
1	32	28,3420
2	15	21,2565
3	13	10,4015
	60	$\sum_{i=1}^{3} e_i \approx 60$

$$J_0 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i} = \frac{(32-28,3420)^2}{28,3420} + \frac{(15-21,2565)^2}{21,2565} + \frac{(13-10,4015)^2}{10,4015} = 2,9628$$

P3) **Establecer nivel de significancia:** $\alpha = 0,05$.

P4) **Región de rechazo de** H_0 **:**

$m = 3,\ \ k = 1$

$$H_0 \text{ v/s } H_1 \ \Rightarrow\ R = \left(\chi^2_{1-\alpha,m-k-1}, \infty\right) = \left(\chi^2_{0,95;1}, \infty\right) = \left(3,84; \infty\right).$$

P5) **Decisión:** Si $J_0 = 2,9628 \notin R = \left(3,84; \infty\right)$, entonces H_0 no se rechaza.

P6) **Conclusión:** La distribución del número de defectos por artículo tiene una distribución de Poisson con un 95% de confianza.

Prueba de Independencia

Supongamos que se tiene una muestra de n datos bidimensionales de las variables X e Y y se clasifica en m categorías $A_1, A_2, .., A_m$ para X y k categorías $B_1, B_2, .., B_k$ para la variable Y, se presenta en la siguiente tabla de frecuencias conjunta:

X / Y	B_1	B_2	…	B_j	…	B_k	Total
A_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1k}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2k}	$n_{2\bullet}$
:	:	:	…	:	…	:	:
A_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ik}	$n_{i\bullet}$
:	:	:	…	:	…	:	:
A_m	n_{m1}	n_{m2}	…	n_{mj}	…	n_{mk}	$n_{m\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet k}$	n

Nota:

- Se cumple que $n = \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{m} n_{i\bullet} = \sum_{j=1}^{k} n_{\bullet j}$
- Cuando las dos características a estudiar corresponden a datos cualitativos se habla de Tablas de Asociación. Además es posible realizar tabulaciones mixtas.

P1) **Plantear Hipótesis:**

H_0 : Las variables X e Y son independientes v/s H_1 : Existe alguna relación entre X e Y.

La hipótesis H_0 es equivalente a probar que $P(A_i \cap B_j) = P(A_i) \cdot P(B_j), 1 \le i \le m$ y $1 \le j \le k$, es decir que los sucesos A_i y B_j son independientes.

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(m-1)(k-1)}$, donde:

n_{ij} : Frecuencia observada de la categoría $A_i \cap B_j$.

$e_{ij} = n \cdot P(A_i \cap B_j) = \frac{n_{i\bullet} n_{\bullet j}}{n}$, es la frecuencia esperada de la categoría $A_i \cap B_j$.

k : Número de categorías para la variable Y.

m : Número de categorías para la variable X.

P3) **Establecer nivel de significancia:** α.

P4) **Región de rechazo de** H_0 **:** H_0 v/s $H_1 \ \Rightarrow\ R = \left(\chi^2_{1-\alpha,(m-1)(k-1)}, \infty\right)$.

P5) **Decisión:** Si $J_0 \in R$, entonces H_0 se rechaza al nivel de significancia α.

P6) **Conclusión:** Se debe interpretar la decisión tomada en P5).

Grado de relación: Para medir el grado de relación entre las variables cualitativas se usa como indicador el coeficiente de contingencia (CC), definido por:

$$CC = \left(\sqrt{\frac{J_0}{J_0 + n}} \right) \cdot 100$$

Ejercicio

Una empresa minera hizo un estudio para verificar si el tipo de trabajo se relaciona con el grado de silicosis de los trabajadores. Para lo cual se elige una muestra aleatoria de 300 trabajadores y se clasifican en la tabla siguiente:

Tipo de trabajo	Grado Silicosis			Total
	I	II	III	
Oficina	42	24	30	96
Terreno	54	78	72	204
Total	96	102	102	300

a) Probar la hipótesis de que el tipo de trabajo afecta el grado de silicosis del trabajador con un nivel de significación de 5%.
b) Determine el grado de relación.

Respuesta

Sea X : "Tipo de trabajo" y Y : "Grado de silicosis".

P1) **Plantear Hipótesis:** H_0 : El grado de silicosis es independiente del trabajo v/s H_1 : Existe alguna relación entre grado de silicosis y el tipo de trabajo.

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(m-1)(k-1)}$

i	n_{ij}	e_{ij}	$\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$
1	42	30,72	4,1419
2	24	32,64	2,2871
3	30	32,64	0,2135
4	54	65,68	1,9494
5	78	69,36	1,0763
6	72	69,36	0,1005
		J_0	9,7683

P3) **Establecer nivel de significancia:** $\alpha = 0,05$.

P4) **Región de rechazo de H_0:**

$$H_0 \text{ v/s } H_1 \Rightarrow R = \left(\chi^2_{1-\alpha,(m-1)(k-1)}, \infty \right) = \left(\chi^2_{0,95;2}, \infty \right) = (5,99; \infty).$$

P5) **Decisión:** Si $J_0 = 9,7683 \in R = (5,99; \infty)$, entonces H_0 se rechaza al nivel de significancia de 5%.

P6) **Conclusión:** Con 95% de confianza podemos decir que existe alguna relación entre grado de silicosis y tipo de trabajo.

Prueba de Homogeneidad

Consideremos k poblaciones independientes, cada una particionada en las clases $A_1, A_2, .., A_m$.

Para cada A_i se definen las probabilidades

$P(A_i \mid población \quad j) = \frac{n_{ij}}{n_j}$, $1 \le i \le m$ y $1 \le j \le k$,

La hipótesis a probar es si cada clase A_i tiene la misma probabilidad en todas las poblaciones. Para realizar la prueba se toma una muestra de tamaño n_j de la población j y se clasifican según la siguiente tabla:

Clases	Población						Total
	1	2	…	j	…	k	
A_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1r}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2r}	$n_{2\bullet}$
:	:	:	…	:	…	:	:
A_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ir}	$n_{i\bullet}$
:	:	:	…	:	…	:	:
A_m	n_{m1}	n_{m2}	…	n_{mj}	…	n_{mr}	$n_{m\bullet}$
Total	n_1	n_2	…	n_j	…	n_k	n

P1) **Plantear Hipótesis:**

$$H_0: \begin{pmatrix} p_{11} \\ p_{21} \\ : \\ : \\ p_{m1} \end{pmatrix} = \begin{pmatrix} p_{12} \\ p_{22} \\ : \\ : \\ p_{m2} \end{pmatrix} = ... = \begin{pmatrix} p_{1k} \\ p_{2k} \\ : \\ : \\ p_{mk} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ : \\ : \\ p_m \end{pmatrix} \qquad \text{v/s} \quad H_1: \text{Existe alguna diferencia.}$$

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(m-1)(k-1)}$, donde:

n_{ij} : Frecuencia observada de la categoría $A_i \cap B_j$.

k : Número de poblaciones.

m : Número de clases.

$e_{ij} = n_j \hat{p}_i$, es la frecuencia esperada de la clase A_i en la población j. Donde se calcula por $\hat{p}_i = \frac{n_{i\bullet}}{n}$ suponiendo H_0 verdadera.

P3) **Establecer nivel de significancia:** α.

P4) **Región de rechazo de H_0:** H_0 v/s $H_1 \Rightarrow R = \left(\chi^2_{1-\alpha,(m-1)(r-1)}, \infty\right)$.

P5) **Decisión:** Si $J_0 \in R$, entonces H_0 se rechaza al nivel de significancia α.

P6) **Conclusión:** Se debe interpretar la decisión tomada en P5).

Ejercicio

En un proceso de fabricación de tornillos, el fabricante quería determinar si la proporción de tornillos defectuosos producidos por tres máquinas variaba de una máquina a otra. Para verificar esto se

seleccionaron muestras de 400 tornillos de la producción de cada máquina y se contó el número de tornillos defectuosos en cada una, obteniendo la siguiente tabla de frecuencias:

Calidad	Máquina			Total
	1	2	3	
Defectuoso	16	24	9	49
No Defectuoso	384	376	391	1151
Total	400	400	400	1200

Realizando la prueba de hipótesis adecuada, verifique si la proporción de tornillos defectuosos no varía entre las diferentes máquinas. Use un nivel de significación de 0.05.

Respuesta

P1) **Plantear Hipótesis:**

$$H_0: \begin{pmatrix} p_{11} \\ p_{21} \end{pmatrix} = \begin{pmatrix} p_{21} \\ p_{22} \end{pmatrix} = \begin{pmatrix} p_{31} \\ p_{32} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \text{ v/s } \quad H_1: \text{ Existe alguna diferencia.}$$

P2) **Estadístico de prueba:** $J_0 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(m-1)(k-1)}$,

i	n_{ij}	$e_{ij} = n_j \hat{p}_i$	$\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$
1	16	16,333	0,007
2	384	383,667	0,000
3	24	16,333	3,599
4	376	383,667	0,153
5	9	16,333	3,293
6	391	383,667	0,140
		J_0	7,192

P3) **Establecer nivel de significancia:** $\alpha = 0,05$.

P4) **Región de rechazo de H_0:**

$$H_0 \text{ v/s } H_1 \Rightarrow R = \left(\chi^2_{1-\alpha,(m-1)(k-1)}, \infty\right) = \left(\chi^2_{0,95;2}, \infty\right) = (5,99; \infty) .$$

P5) **Decisión:** Si $J_0 = 7,192 \in R = (5,99; \infty)$, entonces H_0 se rechaza al nivel de significancia de 5%.

P6) **Conclusión:** Con 95% de confianza podemos decir que existe alguna diferencia entre las proporciones.